15400 Calhoun Drive, Suite 400
Rockville, Maryland, 20855
(301) 294-5200
http://www.i-a-i.com

Intelligent Automation Incorporated

Information Tailoring Enhancements for Large-Scale Social Data

Progress Report No. 2

Reporting Period: December 16, 2015 – March 15, 2016

Contract No. N00014-15-P-5138

Sponsored by
ONR, Arlington VA
COTR/TPOC: Dr. Rebecca Goolsby

Prepared by
Onur Savas, Ph.D.

DISTRIBUTION A

Approved for public release; distribution is unlimited.

Progress Report No. 2

Information Tailoring Enhancements for Large-Scale Social Data

Submitted in accordance with requirements of
Contract #N00014-15-P-5138

Performance period: December 16, 2015 to March 15, 2016
(PI: Dr. Onur Savas, 301.294.4241, osavas@i-a-i.com)

1 Work Performed within This Reporting Period

In this reporting period, we performed the following tasks.

- **Implemented Temporal Analysis Algorithms for Advanced Analytics in Scraawl.** We implemented our backend web service design for the temporal analysis and we created a prototyope GUI web service of Scraawl analytics dashboard.

- **Upgraded Scraawl computational framework to increase robustness.** We imporoved the (i) messaging architecture, (ii) data redundancy, and (iii) service availability of Scraawl computational framework.

- **Delivered Scraawl Version 1.12.2.**

1.1 Implemented Temporal Analytics

In our recent work in Scraawl, we implemented the capability to discover influential users from Twitter. In particular, we developed and implemented the capability to identify influential users using an interaction graph that we build from the collected social media data using Scraawl as shown in Figure 1.

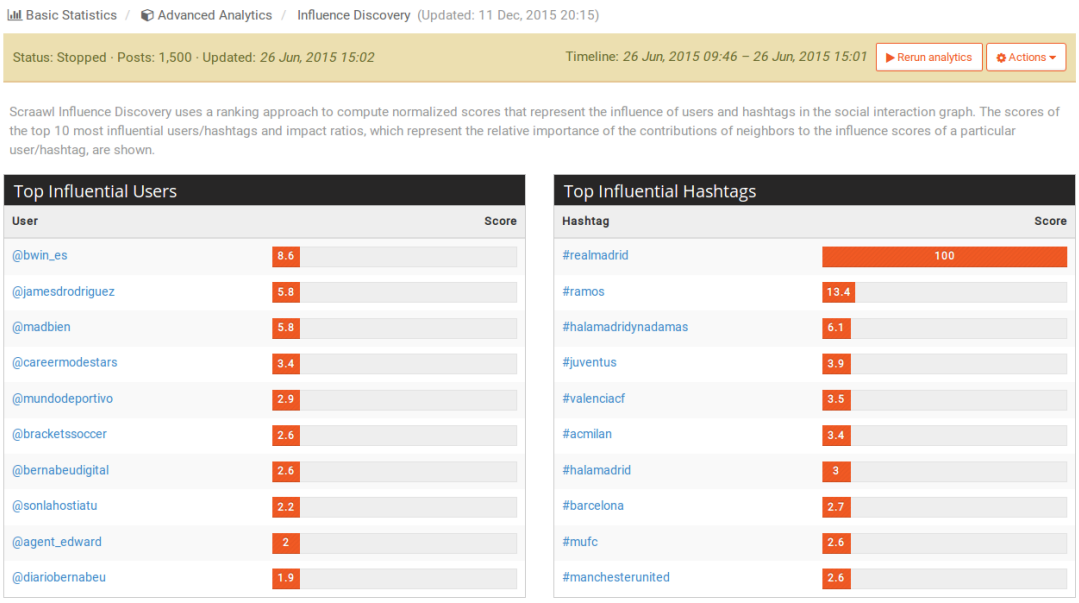

Figure 1: A sample of Influential users results for a Scraawl report.

In this reporting period, we implemented a smart data binning for influential nodes analytical capabilities for twitter data source to evaluate the temporal evolution of this analytical capability.

Our lightweight temporal software implementation is described as follows:

1. We extended the influential nodes web service to have the time period (*IntervalInMinutes*) for a bin in minutes as a parameter. The caller of the web service can run the analytic if he provides the value zero to the *IntervalInMinutes*, otherwise the caller will run the temporal analytics.
2. We modified the tweets database extraction routine to pull the timestamps of the time of tweet creation for every tweet as reported by twitter.
3. We developed a new smart binning routine that has as input the *StartingTime*, *EndingTime*, and an *IntervalInMin*:
    a. We retrieve *hour* and *minutes* from the timestamp of the first tweet in the report.
    b. We calculate the next *EndingTime* of the bin by finding the next minute *m* that satisfy the following equation $m = K * IntervalInMinutes$ among all the minutes in the first hour of the social media data report.
    c. We compute all the bins by shifting the *StartingTime* and *EndingTime* of the first bin *IntervalInMinutes* until the *StartingTime* is greater than the last tweet time in the report
    d. We replace the *EndingTime* of the last bin with the last time of a tweet in the report
4. The posts are divided into batches and grouped by *report_id* and an identical time period using a smart data binning model approach. We developed a routine to return all the tweets in each bin such as the time t of a tweet satisfy the following

condition: *Starting_Time_Of_Bin <= t < End_Time_Of_Bin*

5. We developed a new routine to store the application results along with additional fields = { *StartingTime* of each bin, *EndingTime* of each bin, and the *BinNumber*} in a specific table of MySQL database. We extended the analytics influence results table in MySQL database and the Influence model to address the storage of the additional fields.
6. We run our influence discovery application on those batches of social media data within the report and we store the results in MySql database.
7. On the GUI, we developed a scrolled list of temporal batches for the user to see the evolution of the analytical results.

Example web service request for temporal influential nodes:

```
{
  "topk": "10",
  "topn": "10",
  "IntervalInMin":"15",
  "promise": "http://www.google.com",
  "error": "http://www.google.com"
}
```

Example web service response for temporal influential nodes:

```
{
"status": "OK",
"message": "Analytics completed"
}
```

Example web service request for influential nodes on all tweets:

```
{
  "topk": "10",
  "topn": "10",
  "IntervalInMin":"0",
  "promise": "http://www.google.com",
  "error": "http://www.google.com"
}
```

1.2 Upgraded Scraawl computational framework to increase robustness

In this task we improved the computational framework of Scraawl to make it more robust and handle higher data loads. Figure 2 shows the different components that are part of the Scraawl computational framework, and the section of the framework that was upgraded is highlighted in the figure. As part of this task we focused on improving the (i) messaging architecture, (ii) data redundancy, and (iii) service availability.

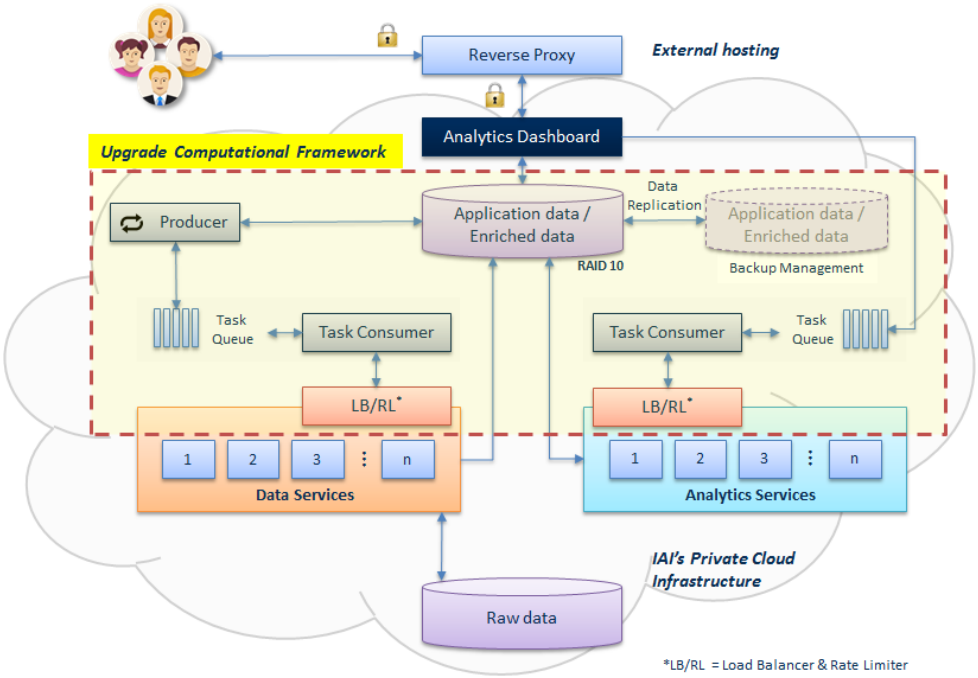

Figure 2: Scraawl computational framework.

In particular, we improved the architecture that handles the *tasking and messaging* aspect of the computational framework. This involved making the message queues *highly available*, providing improved task delegation intelligence, and improved monitoring of the message queues. For *data redundancy* we added more hardware to support replicated data storage. We also improved the availability of the application by providing redundant standby data nodes. To improve the *service availability*, we improved the monitoring intelligence of the computational nodes. We introduced redundant stand by computational nodes that can be added to the production system to handle unexpected service failures. This reduced the downtime and added robustness to the computationally heavy analytic services.

2 Current Problems

None.

3 Work to be Performed in the Next Reporting Period

In the next report period, we will focus on the following tasks:

- We will enhance current Named Entity Recognition (NER) algorithm with additional named entities in Scraawl.
- We will deliver Scraawl 3.

4 Financial Status

Financially, we are in good shape.